Review on Present State-of-the-Art of Secure and Privacy Preserving Data Mining Techniques

J. Pradeep Kumar
Assistant Professor,
Dept of CSE, Aditya College of Engineering, Madanapalle,
jayampradeepkumar@gmail.com

Dr. A. Udaya Kumar
Professor,
Hindustan Institute of Technology & Science, Chennai
acoe@hindustanuniv.ac.in

Dr. T. Ravi
Principal
Srinivasa Institute of Engineering & Technology, Chennai
travi675@yahoo.com

Abstract--As people of every walk of life are using Internet for various purposes there is growing evidence of proliferation of sensitive information. Security and privacy of data became an important concern. For this reason privacy preserving data mining (PPDM) has been an active research area. PPDM is a process discovering knowledge from voluminous data while protecting sensitive information. In this paper we explore the present state-of-the-art of secure and privacy preserving data mining algorithms or techniques which will help in real world usage of enterprise applications. The techniques discussed include randomized method, *k*-Anonymity, *l*-Diversity, *t*-Closeness, *m*-Privacy and other PPDM approaches. This paper also focuses on SQL injection attacks and prevention measures. The paper provides research insights into the areas of secure and privacy preserving data mining techniques or algorithms besides presenting gaps in the research that can be used to plan future research.

Index Terms *– Privacy preserving data mining, k-anonymity, l-diversity, m-Privacy*

*****

I. INTRODUCTION

Privacy preserving data mining refers to the extraction of trends or patterns from data sources without disclosing sensitive information [12]. Security of the data while performing mining operations also important. All existing data mining techniques can be used in order to discover actionable knowledge. As data mining became indispensable in large organizations, security concern is also growing. Towards security and privacy it is essential to have secure and privacy preserving data mining. This paper focuses on various algorithms or techniques that for the present state-of-the-art in the area of secure and privacy preserving data mining. Figure 1 captures the essence of privacy preserving data mining. Many privacy preserving data mining techniques came into existence. They include randomization which advocates adding noise to data in order to mask its attributes; k-anonymity and l-diversity models that prevent inferring sensitive information from publicly disclosed records; distributed privacy preservation which protects sensitive data involved in distributed data mining where data is said to be distributed horizontally or vertically; downgrading application effectiveness in terms of query auditing, classifier downgrading and association rule hiding. Other techniques include group based anonymization, distributed privacy-preserving data mining techniques in the presence of semi-honest and malicious adversaries, privacy preservation of application results, and the curse of dimensionality with respect to limitations of privacy [6]. The applications of privacy preserving data mining are widely used in the area of customer transaction analysis, medical data mining, and home land security to name few. The applications include scrub and data fly systems in healthcare domain, bioterrorism applications, credential validations, identity theft prevention systems, web camera surveillance, video-surveillance, watch-list problems, and genomic privacy [6].

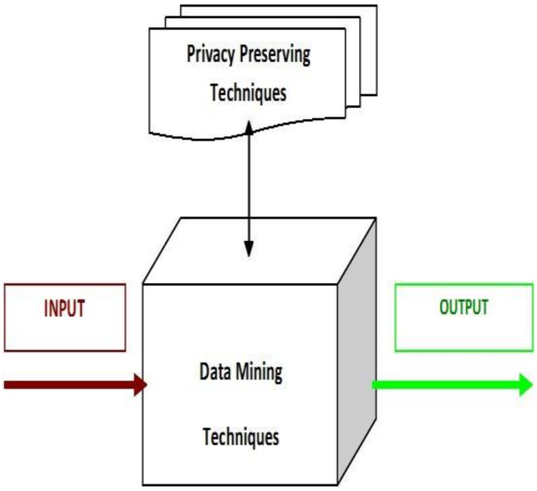


Figure 1 – Schematic flow of privacy preserving data mining

Many researches contributed towards secure and privacy preserving data mining. In [1]-[12] the problem of PPDM was explored. In [2] PPDM is applied for smart homes where security is provided to homes besides preserving sensitive information. In [3] cryptographic techniques are used to achieve PPDM. A protocol named **"private scalar product protocol"** is proposed in [5] for secure mining of massive datasets. In [7] PPDM techniques are applied on vertically partitioned databases. Amplification methodology is presented in [8] for PPDM while standardization technique is employed in [9]. In [11] the Cloud Security Alliance (CSA) presented top 10 big data security and privacy challenges that includes PPDM.

There are many recent contributions towards PPDM and secure data mining as explored in [24]-[38]. In [24] remote data integrity is explored. A novel data leakage prevention method is presented in [25]. Direction and indirect discrimination prevention in data mining was explored in [26] while [27] deals with anonymizatn problem on centralized and distributed data. Incentive compatible PPDM is the main research area in [28] where secure multi-party computations take place for PPDM. In [29] the data mining algorithms such as K-means, SVM classifiers and fuzzy neural networks are used for PPDM. A novel concept such as m-Privacy is introduced in [30] for privacy preserving data publishing. Private information inference attack prevention [31], anonymization for protecting sensitive labels in social networks [32], aggregate knowledge attack prevention [33] , privacy aware data aggregation [34], PPDM on location based queries [35], PPDM on cloud data [36], PPDM for personalized web search [37] and secure mining of association rules [38] are other recent researches in the area of PPDM.

In this paper, we focus on SQL injection attack as well. SQL injection attack is the attack made on web applications to tailor SQL commands and obtain sensitive information. Most of the web applications are vulnerable to this attack. This needs to be taken seriously by application developers, administrators and auditors. Researches in [13]-[23] focused on SQL injection and various prevention approaches. A combinatorial approach is presented in [14] while parse tree is used in [15] for detection and prevention of SQL injection attacks. A novel approach is introduced in [19] by using syntax evaluation and positive tainting. Query tokenization approach is used in [20] for preventing SQL injection attacks. Our contribution in this paper is the review of the present state-of-the-art of Secure and PPDM. The remainder of the paper is structured as follows. Section 2 presents randomization method. Section 3 presents PPDM techniques like *k*-anonymity and *l*-diversity. Section 4 provides information about *t*-Closeness model. Section 5 presents *m*-privacy for collaborative data publishing. Section 6 presents privacy protection in personalized web search. Section 7 presents other PPDM techniques while section 8 presents SQL injection attacks and prevention. Section 9 presents research gaps in the area of secure and privacy preserving data mining. Section 10 concludes paper besides giving directions for future research.

II. Randomization Method

Randomization is a technique that is used to add noise to original data in order to support privacy preserving data mining. Perturbation is used to add noise sufficiently so that values in actual records cannot be recovered. The randomization method can be described as follows. A collection of tuples denoted as X = {x_1, x_2, …, x_n}. For each record noise is added with certain probability distribution denoted as fY(y). The noise components can be denoted as y_1, y_2, …, y_n. After randomization, the distorted records can be denoted as x_1+y_1, x_2+y_2, …, x_n+y_n. The new records thus generated can be denoted as z_1, z_2, …, z_n. As the noise is more the original values cannot be guessed. However, the distortion of the original tuples can be removed and original data can be obtained by authorized person.

The randomization technique is vulnerable to attacks especially when the attacker has prior knowledge. The attacks include **known input-output attack** and **known sample attack**. In case of known input-output attack, the attacker known some records and thus applies linear algebra techniques for reverse-engineering and obtaining to original records i.e. x_1, x_2, …, x_n. The known sample attack on the other hand is launched by an attacker who has many independent distributions of data from which original data is taken and applies a technique known as principal component analysis in order to reconstruct the original data. More details on randomization and attacks can be found in [39] and [40].

III. *K*-Anonymity and *l*-Diversity Models

Public records are used by adversaries to infer unknown information from known data. To overcome this problem *k*-anonymity model came into existence. With *k*-anonymity it is possible to reduce the granularity of representation of data using suppression and generalization techniques. The reduction of granularity is done as much as possible in such a way that any record maps at least *k* other tuples in the dataset. *k*-anonymity is able to protect identity of individuals but unable to protect the sensitive information from being inferred. To overcome this weakness, *l*-diversity came into existence. The technique *l*-diversity maintains diversity of sensitive attributes so as to preserve privacy of data. More information can be found on *k*-anonymity and *l*-diversity in [41] and [42] respectively. *l*-diversity is also known for certain attacks. It does mean that attacker with some background knowledge can launch attacks such as **homogeneity attack** and **background knowledge attack**. The first attack is made when many values of a sensitive attribute are same. Though the data is anonymized, the value of that data which is homogenous in nature can be guessed exactly. The background knowledge attack is made using one or more quasi-identifier attributes and finding association of them with a sensitive attribute. This will allow narrowing down the possible values so as to identify sensitive value. Both attacks are used to predict unknown values from known values.

IV. *t*-Closeness Model

The *t*-closeness model is an improved form of *l*-diversity. *l*-diversity treats all values of given attributes in same fashion irrespective of distribution of values in the domain. As the values might be skewed in the real world, this is not the usual case. This will make hindrances while making more diverse representations in the data. This will help an adversary with background knowledge to launch attacks to obtain sensitive information. This problem is overcome in *t*-closeness as it uses a property which can help in making more diverse representations of data in order to increase privacy. “The distance between the distributions of sensitive attribute within anonymized group should not be different from the global distribution by more than a threshold *t*” is the property used by the *t*-closeness model which makes this fundamentally different from *l*-diversity. In order to quantify distance between any two distributions, the measure used in *t*-closeness is known as “Earth Mover Distance Metric”. *t*-closeness is more effective approach when compared with PPDM techniques came prior to it. More information on *t*-closeness can be found in [43].

V. m-Privacy for Collaborative Data Publishing

Goryczka *et al*. [30] presented yet another technique for PPDM. However, it is meant for privacy preserving collaborative data publishing. When multiple parties collaborate there might be internal attacks that make use of local known knowledge to infer unknown data from data contributed by other partners. The notion of m-Privacy is introduced which ensures that anonymized data complies with privacy constraints. Figure 2 presents overview of the distributed data publishing where (b) represents horizontally distributed database across multiple data providers while the (a) represents publicly available information. In (a) providers anonymize data independently and then aggregated which loses potential integrity of data. In (b) collaborative data publishing approach is followed with arrogate and anonymize policy and the usage of secure multi-party computation techniques [30].

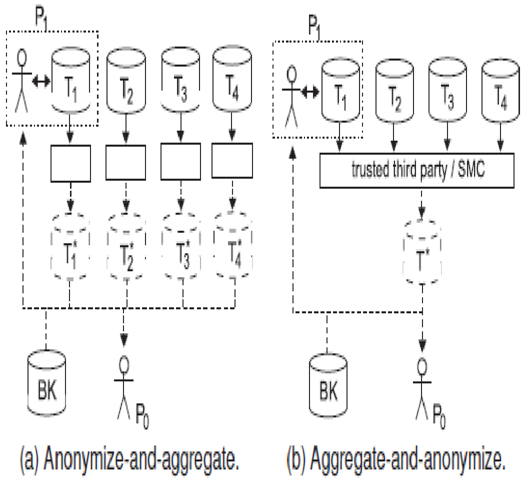

P1
T1
T2
T3
T4
T1*
T2*
T3*
T4*
BK
P0
(a) Anonymize-and-aggregate.
P1
T1
T2
T3
T4
trusted third party / SMC
T*
BK
P0
(b) Aggregate-and-anonymize.

Figure 2 – Overview of distributed data publishing settings [30]

In this *m*-adversary model is built assuming many colluding data providers who can make insider attacks. The provider aware anonymization algorithm with m-Privacy notion privacy preserving collaborative data publishing is achieved [30]. The experimental results reveal that the m-Privacy can be compared with the baseline algorithm as shown in Figure 3.

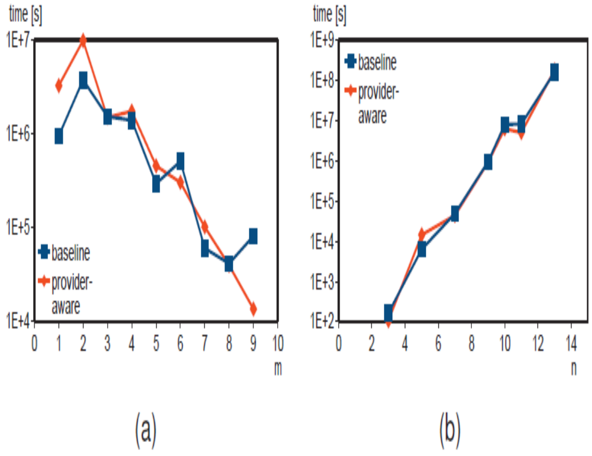


Figure 3 – Computation time against *m* value and number of providers [30]

There is computation time established for both m value and number of data providers. The empirical results reveal that there is almost similar computational time with respect to m value and number of data providers. Another observation is that computational time increases exponentially as *n* value is increased.

VI. Privacy Protection in Personalized Web Search

As personalized web search is growing in popularity, it is essential to have protection to private data in personalized web search. Shou *et al*. [37] studied the protection problem in personalized web search. They proposed a model known as User customizable Privacy-preserving Search (UPS) with two greedy algorithms namely GreedyDP and GreedyIL. The problems with PWS include that there is no support for profiling besides lack in customization of privacy requirements. Iterative use interaction is also missing in some of the techniques. To overcome these drawbacks in [37] the UPS framework is proposed which appears as shown in Figure 4. Online profiler component is responsible to generate user profile based on the privacy preferences selected. Two metrics are considered for generation namely privacy risk and personalization utility.

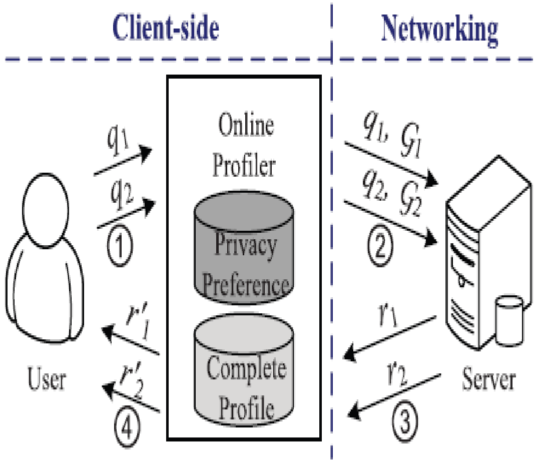


Figure 4 – Overview of UPS framework [37]

When user makes a query both user profile and query are sent to server. Thus the results of the server are personalized and send back to user. The proxy in the client side plays its role in sending queries with profile and receiving personalized responses and presenting them. Two classes of privacy are considered. The first class says that protecting identity of a person is known as privacy. The second one says protection of sensitivity of data is considered privacy. The attack model considered is presented in Figure 5. Eve is able to eaves drop the queries issues by Alice by getting hidden segments thus gaining access to sensitive data. Here the Eve is considered to have assumptions such as knowledge bounded and session bounded. It does mean that no background knowledge about adversary and no perilous session data.

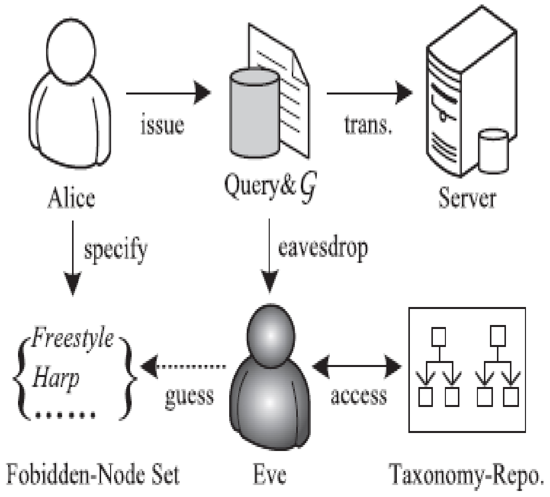
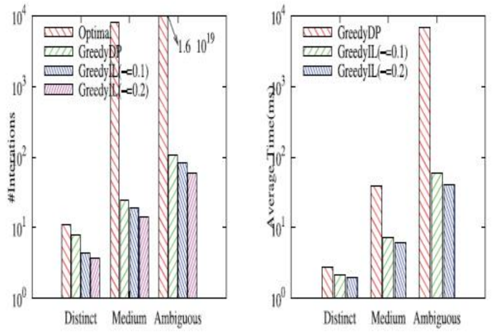


Figure 5 – Attack model of PWS [37]

The generalization techniques and two greedy algorithms contribute towards achieving privacy protection in personalized web search. The efficiency of algorithms is presented in Figure 6.



(a) Results of Average #Iterations (b) Results of Average Response Time

Figure 6 – Efficiency of algorithms [37]

As can be seen in Figure 6, it is evident that different performance is recorded for different algorithms with average number of interactions and average response time. Comparatively GreedyIL has high performance with respect to response time [37].

VII. Other PPDM Approaches

Remote data integrity checking technique is proposed in [24]. They adapted a protocol for public verifiability that is to be done by a third party auditor. The protocol showed good performance in checking integrity of remote data. A misusability weight measure is proposed in [25] where for estimating risk of exposure of data to outsiders. This helped in sanitizing sensitive data in order to achieve PPDM. Discrimination of people and subjecting to unfair treatment based on the data availability is another problem. To solve this problem Hajian *et al*. [26] presented a solution by proposing a methodology that can detect direct and indirect discrimination. In [27] an attempt is made to anonymize distributed and centralized social networks. They presented two variants of anonymization algorithms for achieving this. In [28] incentive compatible privacy preserving data analysis is presented which uses collaborative data analysis technique in which incentives are given to providers of data when they bestow genuine data. Besides, they use Deterministic Non Cooperative Computing model in order to achieve PPDM in distributed environment. In [29] SVM classifiers, neural networks, and K-means algorithms are used for intrusion detection systems. In [31] private inference attacks are studied on social networking and presented mechanisms to prevent such attacks. Sanitization techniques are used to achieve this. In [32] *k*-degree-*l*-diversity technique is used to protect sensitive labels that are involved in social networking. In publication scenarios, a technique is presented by Gkountouna *et al.* [33] for protection from aggregate knowledge attacks. Similar kind of solution for mobile sensing applications is proposed in [34]. Privacy preserving location based queries with protection to content is studied and a solution is provided in [35]. Similar kind of solution is provided for outsourced data in [36]. In [38] secure mining of association rules is presented. They presented an algorithm that is better than Fast Distributed Mining (FDM) algorithm which is a distributed version of Apriori. Two secure multi-party algorithms are proposed in [38] for secure mining of association rules.

VIII. Preventing SQL Injection Attacks

Researches in [13]-[23] focused on SQL Injection attacks and various approaches to prevent them. However, in this section the approach presented by Halfond *et al.* [19] is reviewed. Serious threat is there with SQL injection attacks on web applications. As explored in [19] there are four kinds of SQL injection attacks are identified against Oracle databases. They include Buffer Overflows, Function Call Injection, Code Injection and SQL Manipulation. Out of them the SQL manipulation and code injection are frequently occurring problems and widely known to developer community. SQL manipulation attack is made by modifying SQL query so as to get sensitive information. It is achieved by using UNION and WHERE clauses to return to bypass authentication and get desired data. Code injection attack involves insertion of new code by attacker. One of the best examples of this attack is appending query to SQL SERVER EXECUTE command. These two attacks can occur to many databases while the function call injection and buffer overflows are specific to Oracle databases. Function call injection and buffer overflows occur when attacker injects Oracle database functions into vulnerable SQL commands. To prevent SQL injection attacks an automated solution is presented in [19]. The overview of the approach is presented in Figure 7.

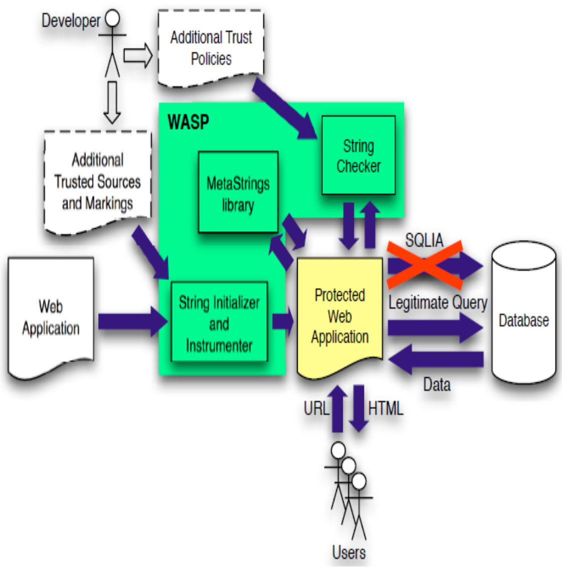

Developer
Additional Trust Policies
WASP
String Checker
Additional Trusted Sources and Markings
MetaStrings library
SQLIA
Web Application
String Initializer and Instrumenter
Protected Web Application
Legitimate Query
Database
Data
URL
HTML
Users

Figure 7 – Overview of Web Application SQL Injection Preventer (WASP).

This solution is made for Java based web applications. Here developer makes use of additional trust policies, additional trusted sources and markings besides a web application. The WASP has components such as String Checker, MetaStrings Library, String Initializer and Instrumenter. With the help of all these components, the protected web application only can send legitimate query to database server. MetaStrings library extends Java's String functionalities. String Initializer and Instrumenter are responsible for identifying and marking trusted strings. Strings are categorized into hard-coded strings, strings that are automatically created by Java and strings originated by external sources.

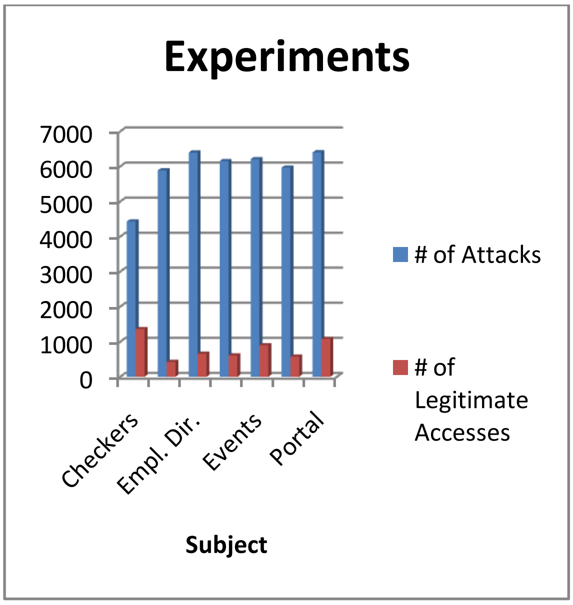


Figure 8 – Experiments made on SQL Injection attacks

As seen in Figure 8, it is evident that number of attacks is made on various subjects. Number of legitimate accesses is also recorded. The number of successful attacks is 0 and the false positives for all subjects are 0. That is the reason the number of successful attacks and the number of false positives are not presented in the graph.

IX. Research Gaps

In [26] the research can be enhanced by analyzing the relationship between privacy preservation and discrimination prevention in data mining. As understood in [27] there is need for distributed versions of *k*-anonymity techniques. In [28] secure multi-party computation techniques can be tailored towards data analysis in DNCC model. In [25] misusability measure can be improved to support multiple publications with Di>1 with different sensitivity combinations. The *m*-Privacy model of [30] can be enhanced by studying a model that addresses the data knowledge of providers of data. The prevention of private information inference attacks presented in [31] can be enhanced by analyzing the information leakage dynamics by altering or removing nodes from graph structure. The solution provided in [32] for protecting sensitive labels can be explored in distributed environment. PWS in [37] can be enhanced further by studying techniques for restricting adversaries with strong background knowledge from to capture series of queries.

X. Conclusions and Future Work

In this paper we review the present state-of-the-art of secure and privacy preserving data mining algorithms or techniques. Various algorithms for privacy preserving data analysis are presented. They include *k*-anonymity, *l*-diversity, *t*-closeness and *m*-privacy besides other PPDM techniques. Other PPDM techniques are presented for anonymization, privacy protection in PWS, incentive compatible PPDM, discrimination prevention, misusability measure, prevention of aggregate knowledge attacks, content and query protection in location based queries, secure processing of multi-keyword ranked query in cloud computing, and secure mining of association rules. We also review the SQL injection attacks and prevention measures. The research gaps are identified for future research. We intend to continue our research in future on secure and privacy preserving data mining approaches and tools.

REFERENCES

[1] Seema Keda, Sneha Dhawale, Wankhade Vaibhav , Pavan Kadam , Siddharth Wani and Pavan Ingale. (April 2013). Privacy Preserving Data Mining. *IJARCCE*. (n.d), p1-4.

[2] Antorweep Chakravorty, Tomasz Wlodarczyk and Chunming Rong. (2013). Privacy Preserving. *IEEE*. (n.d), p23-27.

[3] P.Kamakshi , Dr.A.Vinaya Babu. (April 2010). Preserving Privacy and Sharing the Data in Distributed Environment using Cryptographic on Perturbed data. *JOURNALOFCOMPUTING*.(n.d), p115-119.

[4] K. Srinivasa Rao & B. Srinivasa Rao. (july-august2013). An Insight in to Privacy Preserving Methods. *The Standard International Journals*. (n.d), p100-104.

[5] Bart Goethals, Sven Laur, Helger Lipmaa, and Taneli Mielik¨ainen (n.d). On Private ScalarProduct Computation for Privacy-Preserving.(n.d),p1-17.

[6] Charu C. Aggarwal,Philip S. Yu. (n.d). A General Survey of Privacy-Preserving and Algorithms.(n.d), p12-52.

[7] Cynthia Dwork and Kobbi Nissim. (n.d).Privacy-Preserving on Vertically Partitioned Databases.(n.d), p1-17.

[8] Alexandre Evfimievski,Johannes Gehrke and Ramakrishnan Srikant.(n.d).Limiting Privacy Breaches in Privacy Preserving.(n.d),p1-12.

[9] Stanley R. M. Oliveira and Osmar R. Zaïane. (n.d).Toward Standardization in Privacy-Preserving .(n.d), p1-10.

[10] CHARU C. AGGARWAL,PHILIP S. YU. (n.d). PRIVACY-PRESERVING MODELS AND ALGORITHMS.(n.d), p1-12.

[11] Alvaro Cardenas Mora, Fujitsu Yu Chen, Adam Fuchs, Sqrrl Adrian Lane and Securosis. (November 2012). Top Ten Big and Privacy Challenges. (n.d), p1-11.

[12] Elisa Bertino, Dan Lin, and Wei Jiang.(n.d). A Survey of Quantification of Privacy Preserving.(n.d), p1-20.

[13] Manish Kumar,L.Indu. (2014). Detection and Prevention of SQL Injection attack. ***International Journal** of **Computer Science** and Information Technologies*. 5 (1) (n.d), p374-377.

[14] Dimple D. Raikar, Sharada Kulkarni and Padma Dandannavar. (2012). Preventing SQL Injection Using Combinatorial Approach.*IJARCET*. 1 (n.d), p46-52.

[15] Gregory T. Buehrer, Bruce W. Weide and Paolo A. G. Sivilotti. (n.d). Using Parse Tree Validation to Prevent SQL Injection . (n.d), p1-15.

[16] Dr R.P.Mahapatra and Mrs Subi Khan. (june-2012). A Survey Of Sql Injection Countermeasures. *IJCSES*.(n.d), p55-74.

[17] (n.d). SQL INJECTION **ATTACKS**.(n.d), p149-169.

[18] Zhendong Su,Gary Wassermann.(n.d). The Essence of **Command Injection** Attacks in **Web** Applications.(n.d),p1-11.

[19] William G.J. Halfond, Alessandro Orso, and Panagiotis Manolios.(n.d). Using Positive Tainting and SyntaxAware Evaluation to Counter SQL Injection **Attacks**. (n.d), p1-11.

[20] NTAGWABIRA Lambert,KANG Song Lin.(n.d). Use of Query Tokenization to detect and prevent SQL Injection **Attacks**. *IEEE*.(n.d), p438-440.

[21] PERUMALSAMY RAMASAMY and Dr. SUNITHA ABBURU. (2012). SQL INJECTION ATTACK **DETECTION** AND PREVENTION. *(IJEST)*.(n.d), p1396-1401.

[22] COT **Security Alert** – SQL Injection Attacks.(n.d), p1-2.

[23] Stephen Kost. (2004). An Introduction to SQL Injection Attacks for Oracle Developers. (n.d), p2-24.

[24] Zhuo Hao, Sheng Zhong, Member, and Nenghai Yu. (2011). A Privacy-Preserving **Remote Data** Integrity Checking Protocol with Data Dynamics and Public Verifiability. *IEEE*. VOL.23 (n.d), p1433-1437.

[25] Tamir Tassa and Dror J. Cohen. (2013). Anonymization of Centralized and Distributed Social Networks by Sequential Clustering. *ieee*. 25 (n.d), p311-324.

[26] Murat Kantarcioglu and Wei Jiang. (2013). Incentive Compatible Privacy-Preserving Data Analysis. *ieee*. 25 (n.d), p1324-1335.

[27] Amir Harel, Asaf Shabtai, Lior Rokach, and Yuval Elovici. (2012). M-Score: A Misuseability Weight Measure. *ieee*. 9 (n.d), p414-428.

[28] Sara Hajian and Josep Domingo-Ferrer, Fellow,. (2013). A Methodology for Direct and Indirect Discrimination Prevention in Data Mining. *ieee*. 25 (n.d), p1445-1459.

[29] Slawomir Goryczka, Li Xiong, and Benjamin C. M. Fung. (2013). m-Privacy for Collaborative Data Publishing. *ieee*. (n.d), p1-14.

[30] Raymond Heatherly, Murat Kantarcioglu, and Bhavani Thuraisingham, Fellow. (2013). Preventing Private Information Inference Attacks on Social Networks. *ieee*. 25 (n.d), p1849-1862.

[31] Mingxuan Yuan, Lei Chen, Philip S. Yu, Fellow,. (2013). Protecting Sensitive Labels in Social Network Data Anonymization. *ieee*. 25 (n.d), p633-647.

[32] Russell Paulet, Md. Golam Kaosar, Xun Yi, and Elisa Bertino. (2014). Privacy-Preserving and Content-Protecting Location Based Queries.*ieee*. 26 (.), p1200-1210.

[33] Ning Cao, Member, Cong Wang, Member, Ming Li, Member, Kui Ren, Senior Member and Wenjing Lou. (2014). Privacy-Preserving Multi-Keyword Ranked Search over Encrypted Cloud Data. *IEEE*. VOL. 25, (n.d), p222-233.

[34] Lidan Shou, He Bai, Ke Chen, and Gang Chen. (2014). Supporting Privacy ProtectionPersonalized Web Search. *IEEE*. 26 (n.d), p453-467.

[35] Tamir Tassa. (2014). Secure Mining of Association Rules in Horizontally Distributed Databases. *IEEE*. 26 (n.d),p970-983.

[36] Agrawal R., Srikant R. Privacy-Preserving Data Mining. *Proceedings of the ACM SIGMOD Conference*, 2000.

[37] Agrawal D. Aggarwal C. C. On the Design and Quantification of Privacy-Preserving Data Mining Algorithms. ACM PODS Conference, 2002.

[38] Samarati P.: Protecting Respondents' Identities in Microdata Release. IEEE Trans. Knowl. Data Eng. 13(6): 1010-1027 (2001).

[39] Machanavajjhala A., Gehrke J., Kifer D., and Venkitasubramaniam M.: l-Diversity: Privacy Beyond k-Anonymity. ICDE, 2006.

[40] Li N., Li T., Venkatasubramanian S: t-Closeness: Orivacy beyond k-anonymity and l-diversity. *ICDE Conference*, 2007.